DOCUMENT RESUME

ED 266 174 TM 860 118

AUTHOR Pace, C. Robert; And Others
TITLE The Credibility of Student Self-Reports.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Nov 85
GRANT NIE-G-83-0001
NOTE 64p.; In: "Resource Papers and Technical Reports. Research into Practice Project" (TM 860 116).
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Academic Achievement; Attitude Measures; College Students; *Error of Measurement; Higher Education; Multivariate Analysis; Public Opinion; *Questionnaires; *Reliability; *Self Evaluation (Individuals); Student Characteristics; *Surveys; Tables (Data); *Validity
IDENTIFIERS College Student Experiences Questionnaire; Entering Student Survey; Higher Education Research Institute; *Self Report Measures; Student Information Form

ABSTRACT

This report shows that there are many ways to confirm the accuracy, reliability, and validity of student self-reports. Examples from higher education and from public opinion polls and general surveys demonstrate some of the common sources of measurement errors and errors of substance. Part 1 of the report summarizes a few highlights from the literature, and adds comments from the author's research. Part 2 begins by briefly reporting a tabulation of "missing cases" in three questionnaires for college students. Two of these instruments are the Entering Student Survey, distributed by the American College Testing Program, and the Student Information Form, distributed by the UCLA Higher Education Research Institute; both are widely used, and each has the same general purpose and is intended for the same type of population. Following this, the College Student Experience Questionnaire, designed to be filled out by undergraduates toward the end of the academic year, is discussed in detail. Test-retest comparisons of this questionnaire are used as examples of how subjective responses can be objectively validated. Predictive and construct validity of this questionnaire are examined using multivariate statistical procedures. (LMO)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED266174

THE CREDIBILITY OF STUDENT SELF-REPORTS

Prepared for the Center for the Study of Evaluation

Graduate School of Education, UCLA

by C. Robert Pace

with the assistance of

Doris Barahona
David Kaplan

November 1985

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☑ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C. Griffith

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 860 118

Acknowledgments

I am indebted to Doris Barahona and David Kaplan for important portions of this report. Both are graduate student research assistants in the Graduate School of Education. Following discussions with Doris Barahona about the sort of internal cross-tabs that might bear on the credibility of student reports to the College Student Experiences questionnaire, she identified a whole array of questions and answers that seemed to be relevant, and then obtained the results via many computer printouts. David Kaplan explored the possible relevance of multivariate statistical analyses for judging the predictive and the construct validity of student self-reports in the College Student Experiences questionnaire.

In the process of thinking about and then producing the present document, I have benefited from discussions with these colleagues and I have welcomed and appreciated their interest.

C. Robert Pace
November 1985

INTRODUCTION

Whenever one presents the results of a questionnaire survey, there is always someone who says "But those are only opinions". If the results come from a survey of students, the put-down response is "But those are only students' opinions", as if, coming from students, the results are even less believable. If the comment comes from someone in the "hard" sciences, it is likely to be "But you only have 'soft' data".

It's interesting that this sort of knee-jerk disbelief does not automatically occur in response to other surveys. The Census Bureau conducts many surveys that ask about people's opinions and plans. There are surveys to estimate consumer confidence which are taken seriously by economists and entrepreneurs. Political opinion surveys are carefully studied by candidates for office. Opinion surveys are an important aspect of market research. There is, of course, a certain skepticism about the credibility of some self-reports to the Internal Revenue Service. But on the whole, opinion polls, survey research, and questionnaires are widely accepted methods of inquiry, and certainly a very significant feature of scholarship in the social sciences.

Opinion polls and attitude surveys, like other inquiries, are subject to errors of measurement. For more than fifty years there has accumulated a very large body of research on possible sources of error, and on ways to estimate reliability and validity. The Public Opinion Quarterly regularly publishes scholarly articles on the methodology of polls and surveys. The major polling agencies are especially sensitive about the accuracy and validity of their reports. Some of the best known survey centers are

university-based -- as the National Opinion Research Center at the University of Chicago, and the Institute for Social Research at the University of Michigan.

In higher education, and in education generally, questionnaires are quite common. There has also accumulated over a period of years a body of research on the credibility of students answers to questionnaires. The present report on the credibility of student self-reports is a preliminary document that should, and perhaps may, become a more thorough and scholarly document at some future date. Meanwhile, we aim to present a few highlights from the large literature on measuring attitudes and other subjective phenomena, note some of the accuracy checks that have been made with respect to college student questionnaire responses, and then examine briefly the features of two current questionnaires for entering college students and explore more extensively one current questionnaire for undergraduates to illustrate a variety of reliability and validity estimates that can sometimes be produced to demonstrate the credibility of students answers.

PART 1

ISSUES, ANSWERS, AND ADVICE

The Russell Sage Foundation has recently published a definitive two volume document entitled Surveying Subjective Phenomena, (Turner and Martin, Editors) 1984. For anyone who wishes to review the literature of research on this topic, those two volumes are a fairly complete answer. In addition, the Russell Sage Foundation has also published a book by one of the most highly regarded scholars, Otis Dudley Duncan, Notes on Social Measurement: Historical and Critical, 1984, which deals with the whole domain of counting and classifying demographic and other elements, from antiquity to the present.

In 1976 the College Entrance Examination Board published a monograph by Leonard Baird, Using Self-Reports to Predict Student Performance, which reports much of the evidence from college student surveys about the accuracy of their responses to questionnaire items, as well as their utility for prediction.

Part 1 of this report is not a review of the literature in the usual sense. No attempt is made to cite chapter and verse from dozens of studies. Rather, everything (except as may be subsequently noted) that will be mentioned comes from one or more of the four major sources just cited. What follows, then, is my summary of what I regard as a few highlights from the literature, plus some of my own contributions to that literature over the past 50 years.

Varieties of Self-Reports

Some self-reports merely ask for obvious, easily verifiable information, such as age, sex, marital status. It is a subjective or individual answer to an objective question. At the other end of the spectrum are questions and answers both of which are entirely interpreted by the individual. A good example is the following question: "Taken all together, how would you say things are these days -- would you say that you are very happy, pretty happy, or not too happy?" An example from a survey of college alumni is the following: "What is your present feeling about your college? -- strong attachment to it, pleasantly nostalgic but no strong feeling, more or less neutral, generally negative, thoroughly negative". The meaning of the question and of the response is determined by the respondent, and can be directly known only by the respondent.

In one part of the appendix to Volume 1 of the Russell Sage report there is a "Scheme for classifying survey questions according to their subjective properties" (pages 407-431). The main categories of this scheme illustrate the varieties of self-reports one encounters in surveying subjective phenomena. There are three dimensions. The first is the referent of the question: objective versus subjective events. Objective questions refer to events that can be externally observed. Subjective questions refer to internal conditions, intuitions, beliefs, etc., which are directly knowable only by the individual. The second dimension is the nature of the judgment. Such judgments might involve beliefs, attributions, or valuations, and they involve different intellectual tasks. Simple judgments about the occurrence of events primarily involve

recall. Attributions require generalizations and inference. One finds very generalized referents such as "most people", "all in all", "people running the country today", "most faculty members", etc. The interpretation of answers is complex and surely suggests the importance of skepticism. Valuations include questions about preferences, likes and dislikes, approval ratings, attitudes toward people, groups, organizations, policies, subjective sentiments such as confidence ratings, satisfactions, problems and worries. The third dimension is the object of the report: self versus other. Is the respondent being asked to report about himself? If so, do people tend to present themselves in a good light? How do these self-perceptions influence one's perception of others?

These three broad categories, albeit overlapping in some respects, are useful to keep in mind as one examines the content of questionnaires: the referent of the question, the nature of the judgment, and the object of the report.

Errors of Measurement

In questionnaire surveys of college students the chief source of unrepresentative results are the nature and size of the sample, and the proportion of people who return the questionnaire. Students in a large introductory psychology course are often asked or required to respond to some questionnaire. They, of course, are not a representative sample of anything. For relatively small colleges, the best advice is to give the questionnaire to everyone, thus bypassing the sampling problem. In big universities, the task of having all entering freshmen respond to a

questionnaire is never successfully completed. If one can get two-thirds or three-fourths of the population one is doing rather well. There are good studies that have obtained data from a broad assortment of students and institutions; but nothing comparable to a national public opinion poll in its representativeness. The more significant problem, however, is in the response rate. Whether questionnaires are distributed via the U.S. Postal Sevice, or whether they are put in a campus mailbox, many are never returned.

In a national questionnaire survey of students and alumni which I carried out in 1969, involving random samples at about 75 colleges and universities, the median response rate to the freshman questionnaire was 80%, for the upperclassmen questionnaires the median response rate was 66%, and for the alumni samples the median response rate was 58%. The questionnaires, each about 16 to 20 pages in length, were attractively designed and printed; most colleges used one followup reminder; and for the alumni samples there were two followup reminders.

Even if one had returns from everyone the basic conclusions would not change significantly; but probably in all questionnaire surveys there is some selectivity or bias among those who respond. In the 1969 study the poorest rates of return from freshmen and upperclassmen came from the large institutions; but in the alumni questionnaire the differences in return rates were not related to size, they were related to institutional selectivity and prestige. In the elite categories, only 2 in 20 (10%) had an alumni response rate of less than 50%; in the middle category scholastically, there were 10 of 39 (26%) with a response rate of less than 50%; and in the least selective category, there were 5 of 15 (33%) with

fewer than 50% returns from their alumni.

In two recent questionnaire surveys of UCLA undergraduates, the response rates have been between 45% and 50% There are, of course, ways to increase the rate of return of mailed questionnaires. Unfortunately, for academic researchers, they are very costly and the money is not forthcoming.

Unlike the usual procedure in academic surveys, the national opinion polling agencies collect their data by interviews. The carefully designed stratified area sampling techniques do, in fact, produce reasonably reliable and valid results. The magnitude of non-response is minimal because the interviewers's job is to get everyone who fits the sample specifications.

On several past occasions I have suggested that periodic polls of college students might be very worthwhile. But they would require developing an adequate base for sampling, and this does not now exist. The carefully designed sampling procedures, and the resulting national samples for public opinion polls, are not applicable to the college population.

There are several other aspects to the present topic of measurement error. These relate to the estimation of reliability. Does one get similar answers to the same questions from comparable samples? In a test-retest situation, do people give the same answer the second time that they gave the first time? Do slightly different questions about the same topic result in generally similar responses? Most surveys in social science and in higher education do not report answers to any of these questions, and presumably do not collect evidence about any of these matters. But they should. And at least periodically they have.

In 1948 a 16-page questionnaire was mailed to a sample of Syracuse University alumni. The questionnaire included two types of items which were subsequently readministered to a small sample. The questionnaire contained eleven Activity Scales of eleven items each, labeled Politics, Civic Affairs, Religion, Art, Music, Literature, and Science. The subjects checked each activity they had engaged in during the past year. The scales were Guttman-type scales in that participation in the more dificult activities tended to subsume participation in the easier and more common activities. The score on each scale was simply the number of activities checked. Then there were nine Opinion Scales of six items each, labeled Politics, Civic Relations, Government, the World, Philosophy, Art, Music, Literature, and Science. The statements in the opinion scales were written to reflect basic concepts or generalizations about the topics, generalizations reflecting a consensus of experts in the field, so that it was possible to score each scale simply by counting the number of statements on which one's opinion agreed with the opinions of the experts. Each statement was answered on a five point scale, from Strongly Agree to Strongly Disagree. Six months after the initial sample of 2500 had filled out the questionnaire, a second copy was sent to a small group of 120, receiving 68 in return. The test-retest consistency of scores over this six-month interval was computed. For the Activity Scales, the correlations ranged from .70 to .89, with a median of .83. For the nine Opinion scales the median test-retest correlation was .65, with seven falling between .60 and .70, and two much lower ones of .40 and .31. Consistency of responses was also checked item by item. For the Activity items, the averge percent

of identical responses was 85, with a range from 83 to 87. For the Opinion items the average percent of identical responses was 75, with a range from 68 to 84. The above test-retest data were reported in an article by Pace, "Opinion and Action: A Study in Validity of Attitude Measurement", Educational and Psychological Measurement, Vol. 10, No. 3, 1950, pages 411-429.

The ACT Evaluation/Survey Sevice, Users Guide, 1981, reported test-retest results on ACT's Student Opinion Survey for a group of students at one university who responded to the questionnaire a second time approximately two weeks after the initial response. The average percent of identical responses on the two administrations was 98% for demographic background items (age, race, sex, etc.), 90% for other background items such as hours worked per week, occupational plans, etc., and 93% for items about the usage of college programs and services. For "Satisfaction" items (responses on a five-point scale from Very Satisfied to Very Dissatisfied) referring to such matters as academic aspects of the college environment, rules and regulatioms, facilities, college services, etc., the percent of identical item responses was typically about 64%, and the percent of responses within one scale point of the identical response typically about 95%.

In the American Council on Education Research Report, Vol. 7, No. 2, 1972 by Boruch and Creager, entitled Measurement Error in Social and Educational Survey Research, two examples of test-retest comparisons are cited. One example administered a questionnaire twice, with six weeks intervening, to a group of 107 college students. Questions about students previous achievements resulted in 90% to 100% agreement. Answers to other

facts -- such as father's education and occupation, high school grades, etc., had agreement percentages from 74% to 92%. Attitudinal items, and questions about future plans typically involved agreement in the 60-70% range. The other example was the readministration of the ACE freshman survey questionnaire to 202 students following an interval of two to three weeks. Test-retest correlations for different types of items were as follows: demographic characteristics, mostly .96 to .99; sources of financial support, mostly .85 to .88; self-reported attributes of parents, mostly .60 to .82; items estimating the chances of future events (such as graduating with honors, joining a fraternity or sorority, failing one or more courses, changing career choice, etc.), mostly .58 to .88 with a median of .78; items about life goals such as the importance of being very well-off financially, raising a family, keeping up with political affairs, helping others in difficulty, mostly from .65 to .87 with a median of .73; attitudes toward the importance of various federal actions such as pollution control, school desegregation, veterans benefits, consumer protection, correlations ranging from .41 to .83 with a median of .63; and items about attitudes toward various campus and social issues such as faculty promotions should be based on student evaluations, marijuana should be legalized, with test-retest correlations ranging from .57 to .88 with a median of .66.

Both the ACT and ACE reports show that the greatest variability in responses are found in relation to questions that are ambiguous, or about topics which students may not have given much prior thought or concern , or about attitudes which are themselves subject to various interpretations. In some cases, the test-retest correlations are low enough to raise doubts

about the value of the responses, especially when the test-retest interval is only 2 to 6 weeks. For the more specific items, consistency of responses was quite high.

In public opinion surveys there have been some examples of comparing the results to the same questions when asked by different survey organizations. The closest or most carefully controlled conditions are called tandem surveys. In one such tandem survey, NORC and Roper each drew probability samples and proceded to administer the survey in their customary fashion. This was a survey about public use of and attitudes towards television. Differences in the results were small; but there was a clear effect related to how the organization determined the "don't know" responses. On 52 comparisons, NORC had fewer DKs on 42 items, Roper fewer on 4 items, with no differences on the other items. In another study, a survey about public attitudes and knowledge concerning survey practices, the sample was drawn by the Survey Research Center, and the cases randomly assigned to SRC and Census Bureau interviewers. In general, the results were fairly similar. However the interviewee refusal rate was 6% to the Census Bureau interviewers and 13% to the SRC interviewers.

A summary tabulation reported in Volume 1 of the Russell Sage publication, of 126 instances in which the same questions were asked by different survey measurement programs at about the same time shows that in 45 of the instances there were differences beyond the level typically allowed for sampling error. Such differences could have come from many sources -- context, interviewer effects, training and staff differences, etc. Some of the differences were clearly attributable to how DKs were handled. Variations in practices produce differences in the products; but

most of these differences are relatively small. When the conditions are most comparable, as in tandem surveys, the results are highly congruent.

Errors of Substance

Whether people report accurately about their conditions or their behavior is, in one sense, an error of measurement and in another sense an error of substance. In surveys of college students there is a good deal of evidence that self-reports about their school grades, and about prior accomplishments are very accurate. Much of this literature has been summarized by Leonard Baird in the monograph he wrote for the College Board in 1976. Are student's self-reports of their grades accurate? Baird himself found that the correlation between college-reported and student-reported grades was genrally about .87. In a study of self-reported and transcript-reported grades, by Nichols and Holland in 1963 among National Merit Scholars, cited by Baird, the correlation was .96. Maxey and Ormsby in 1971 reported correlations between self-reported and school-reported grades in a sample of nearly 6000 students in 134 schools to be on the average in the mid eighties. They found that 98% of the students' reported grades were accurate within one grade. Baird concludes from many studies that "research accumulated over 30 years, using various methods, in samples of grade school students, high school students, college applicants, junior college students, four-year college students, and professional school students, adds up to one conclusion: students' reports of their grades are about as useable as school-reported grades". (page 8). Moreover, self-reported grades predict future grades as well as

or better than college entrance tests of academic ability. It seems fair to conclude that, at least for some kinds of questions, errors of substance in the answers are minimal.

The data from the above studies are a good example of what one can expect when the questions are clear and specific, and when the response options are equally clear and specific. Students know the definition of grades and they know their own grades. Consequently, one can have confidence that the subjects can answer the questions. But in many surveys no such clarity is evident.

Evidence from the larger survey research literature also confirms the accuracy of self-reports about various specific conditions or behavior. For example, correlations between employers records about wages, duties, etc., and application blank work histories were generally .90 or greater. Adults reports of whether they owned their home were 96% accurate, had a valid library card 87% accurate. One needs to be reminded here, that "official records" are not always 100% accurate.

Perhaps one of the most serious errors of substance arises from variations in the content, or wording, of the questions, and from the context in which the questions are used. There are some classic examples of this. The following question was asked in a national sample poll: "Do you think the United States should let Communist newspaper reporters from other countries come in here and send back to their papers the news as they see it?" Half the questionnaires asked this question after another question on whether the Soviet Union should allow in American newspaper reporters; and the other half of the questionnaires asked the questions in the reverse order. When the question about communist reporters was asked first, 55% of the people agreed; but when the question about American

reporters was asked first, 75% agreed. Or, consider the following two questions: 1) Do you approve or disapprove of a married woman earning money in business or industry if she has a husband capable of supporting her? (65% of a national sample approved); 2) If there is a limited number of jobs, do you approve or disapprove of a married woman earning money in business or industry when her husband is able to support her? (Only 36% approved!) Here is another example of different answers from slight differences in wording. "Do you think the United States should forbid public speeches against democrary?" (Yes, 54%.) Do you think the United States should allow public speeches against democracy?" (No, 75%).

Another type of error, potentially causing substantive or interpretive difficlties, is the use of response options that each person interprets in his own way. Examples of such response options are the use of words or phrases such as frequently, occasionally, rarely, most of the time, very much, quite a bit, usually, seldom, a great deal, very little, etc.. Presumably words such as always and never mean the same to everyone. But how often is "often"? And how much is "very much"?

Pace and Friedlander, "The meaning of response categories: how often is occasionally, often , and very often?", Research in Higher Education, Vol. 17, No. 3, 1983, addressed this issue using data from the College Student Experiences questionnaire. Participation in various college activities were initially indicated by the responses "never", "occasionally", "often" or "very often". Later in the questionnaire seven of the same activities were responded to as follows: For each of the items below, fill in one of the spaces to the left which best indicates the number of times you have engaged in the activity. These more specific

responses were: "never", "once or twice during the year", "about three to six times during the year", "about once or twice a month", "about once a week" and "more than once a week". By this means we were able to show what students meant (number of times) by the more general words. The results, as one would expect, revealed considerable overlap by what was meant by occasionally, often, and very often. But there was also a clear concentration or clustering of responses as one moved from occasional to often, and from often to very often. The meaning or definition of these general descriptors was different, depending upon the topic; but within the same topic the differences between colleges or types of students were quite small. In general, the definition of "occasionally" at one college was similar to its definition at other colleges, given the same topic.

Every respondent knows perfectly well that "very often" is more than "often", and that "often" is more than "occasionally". Thus, the direction of the scale is recognized by everyone. But the specific meaning attached to the labels is an individual judgment. There were few obviously implausible responses -- such as students who initially said "occasionally" or "often" but later said "never"; or students who initially said "occasionally" but later said "more than once a week". These discrepancies constituted from 2% to 10% of the total responses.

Comparative judgments of this sort necessarily reflect some reference group in the mind of the judges. On this questionnaire, we assume that the college peer group is the reference group, and that the answers reflect an awareness of what is customary in one's own behavior and in the behavior of the peer group.

The point of these observations about the subjective definition of response choices is that one should get, if at all possible, some sort of evidence about what people mean by their choices. This same advice applies to opinion polls which ask about degrees of happiness, satisfactions, confidence, or other subjectively defined responses.

PART 2

THREE COLLEGE QUESTIONNAIRES

Efforts to evaluate the influences of college on students' learning and development should draw upon many sources of evidence. For much of this relevant evidence the students themselves are the source; and the most common method for obtaining that evidence is a questionnaire.

Here, for example, are four crucial questions.

1. Who goes? What do we know about the entering students: their high school record and test scores, their family background, financial status, their interests, expectations, aspirations, past achievements, etc.? Some of this information can be obtained from records, but some can be obtained only by asking the students themselves.

2. What do they do after they get there? Some answers can be obtained from college records -- such as, campus residence and major field, but for other sorts of behavior -- such as the time and effort devoted to study, contacts with faculty, involvement in extra-curricular activities, use of the library, etc. -- the answers can only come from students' responses to questionnaires.

3. What's it like? Physical facts -- such as big school, small school, and big city, small town -- are important. So also are students' perceptions of the campus environment or atmosphere. What is stressed? What is expected? How do people relate to one another -- friendly, supportive, or not? Answers to these questions can only come from the students themselves.

4. What do they get out of it? Knowledge, basic skills, and abilities relevant to a career, relevant to personal maturity and life

satisfaction, relevant to civic enlightenment -- these are some of the possible and intended results. Achievement tests, abilty tests, personality tests, etc. can provide some of the answers. It may also be important to find out what the students themselves think they got out of college; and here again one relies on responses to questionnaires.

Questionnaires can, and I think should, be regarded as a form of test or measuring instrument. Many questionnaires are in fact regarded as tests by those who construct them. So, we have tests of attitudes and beliefs, vocational interests, pesonality traits, etc.. A variable or dimension to be measured is defined, sets of items are developed to measure it, and the reliability and validity of the results are determined. The process is similar to the construction of an objective achievement test, or a test of developed abilities such as the Scholastic Aptitude Test. Attitudes, interests, beliefs, etc., are subjective phenomena. The answers one gives to a question about interests or opinions is determined by the individual. The student decides whether he agrees or disagrees with some statement, or likes or dislikes some activity, or person, or condition. The good published tests of personality, interests, or values provide extensive data regarding their reliability and validity -- tests such as the Minnesota Multiphasic, the Omnibus Personality Inventory, Holland's Vocational Preferences Inventory, the Allport-Vernon-Lindzey Study of Values. In some tests of this sort, the authors have included a few items to detect whether a student is giving false or improbable answers -- a practice which recognizes the importance of estimating the credibility of self-reports.

Many of the questionnaires used in studies of higher education are not designed as tests in the classical sense. They consist of sets of items,

often grouped or classified under certain topics, but having no underlying or scorable dimension. One finds for example, various items about students use of counseling services, or various items about students opinions of teaching practices, or various items about students attitudes regarding political and economic issues. The items are no doubt regarded as interesting and the answers useful to know. But the content is best described as a classified catalogue rather than as a theoretically or conceptually based set of dimensions or characteristics. The value of the question and the credibility of the answer has to be examined item by item. There is nothing inherently unreliable or invalid about a one-item test. Most public opinion polls are really one-item tests. But it is important to realize that variations in responses are often caused by variations in the phrasing of the question. Slight changes in wording can produce significant changes in responses. Consequently, the meaning of the answers rests on a slender base.

To begin Part 2 we briefly report a tabulation of "missing cases" in three questionnaires for college students. The results illustrate some of the principles and advice given in Part 1, and serve to confirm, with these three current cases, the merit of that advice. Then, the main content of Part 2 is a detailed examination of one questionnaire to illustrate some of the internal and external checks that can be made to assess the reliability and validity of students responses. The content of this one instrument -- Pace's College Student Experiences Questionnaire -- makes meaningful cross checks possible, for it bears upon all four of the topics noted in the introduction to Part 2: Who goes? What do they do after they get there? What's it like? and What do they get out of it?

Missing Cases: What types of questions are not answered?

To provide some current illustrations of non-response to questionnaire items we have examined two widely used instruments, each having the same general purpose and each intended for the same type of population. The first is the Entering Student Survey, distributed by the American College Testing Program. The second is the Student Information Form, distributed by the UCLA Higher Education Research Institute.

Both of these questionnaires are introduced with assurances regarding the confidentiality of the students' responses. The HERI questionnaire says "Identifying information has been requested in order to make subsequent mail followup studies possible. Your response will be held in the strictest professional confidence". The ACT questionnaire says, "The information you supply on this questionnaire will be kept completely confidential. Your name, address, and Social Security number will enable college officials to identify your responses and to contact you directly. The data you supply will be used for research purposes and will not be individually listed on any report. If, however, any question requests information you do not wish to provide, feel free to omit it."

Both questionnaires have many similar and in some cases identical items, for example: age, race, sex, marital status, planned college residence, high school grades, planned college major, planned occupational choice, sources of funding, reasons for going to college. Straightforward identification questions, and questions about specific activities, reasons for going to college, etc. are typically omitted by fewer than 4% of the cases, and often by fewer than 2%. The questions which are omitted by the

largest percentages of respondents are ones related to money, religion, expected major field and occupation, and assorted items about personal and social values.

On the HERI questionnaire there are typically about 12% to 13% who do not answer the items about parents income, and sources of funding for college. Many of those items identify specific dollar amounts -- parents total income -- or a specific fact -- listed as a dependent on federal income tax return. No doubt in some instances the students do not know the answers; and perhaps in other instances they regard the question as inappropriate. The ACT does not ask about dollars; it asks whether various sources of funding are a major source, minor source, or not a source. Eleven sources are listed, and about 5 1/2% to 9% of the students do not respond.

The HERI questionnaire asks the students to indicate the religious preference of self, father, and mother. From 15% to 17% do not answer the question.

On the HERI questionnaire 6% of the entering freshmen do not indicate their probable undergraduate major, and nearly 7% do not indicate their probable career occupation. On tne ACT questionnaire the percent of omits is 12% for the probable major and 16% for the probable occupation. The reason for these larger numbers may be owing to the format. The ACT survey has a separate sheet inserted with the questionnaire listing many major fields and occupations. The student finds the 3-digit code that best describes his plans and then fills in these numbers on the questionnaire. Apparently some students just don't bother to do this. On the HERI questionnaire the various fields and occupations are listed on the

quesionnaire itself, making the response easier to record. In both cases, however, it seems reasonable to suspect that asking entering freshmen about their probable college major and their probable occupation is not viewed as an answerable question by some students. In fact, on a different part of the HERI questionnaire more than 20% of the students said the chances were very good that they would change their major and change their occupational choice.

In both questionnaires, items about such topics as reasons for going to college and reasons for going to this particular college, were omitted by only 2% to 4% of the respondents in most instances. The ACT questionnaire has a section labeled "college impressions" where students are asked to indicate their agreement with various statements about the college environment -- such as "students at this college are friendly", "this college offers many cultural events and programs". Typically about 3% omit these items; although one wonders about the basis for the answers because often one's impressions, in advance of actual experience, reflect common stereotypes about what college is like.

The HERI questionnaire asks students questions about various political, social, and educational policies -- such as "abortion should be legalized", "college grades should be abolished", "the federal government is not doing enough to control environmental pollution". Typically about 5% to 8% of the students do not answer these questions. Another question asks students to characterize their political views, as far left, liberal, middle-of-the-road, conservative, or far right. About 5% do not answer the question.

For all of the above data, the information about the ACT questionnaire comes from a normative report based on about 16,000 cases in which the number of "blank" responses to every item is listed. For the HERI questionnaire the data come from the 1983 report of freshmen norms in which the data for one sample college are shown, having about 2,300 cases. The complete normative report does not show missing cases.

Except for the questions about major field and probable occupation, the number of "omits" in the ACT questionnaire is generally smaller than in the HERI questionnaire. There may be several factors accounting for this. The ACT questionnaire is shorter. The format and organization are also clearer. Section 1 is labeled Background Information, Section 2 is Educational Plans and Preferences, Section 3 is College Impressions. Although in some parts the print is quite small, each part is enclosed in a box, with the question or topic itself in boldface capital letters. Perhaps more important is the likelihood that most students would not view any of the questions as offensive or intrusive. There is no invasion of privacy of the sort that might influence one to omit the answer or to give a socially desirable answer rather than a more forthright answer. One can easily regard the questions as appropriate to ask of entering students because of the educational relevance of the questions.

The HERI questionnaire, although of the same four-page length as the ACT, has many more items, and the format consequently appears crowded. Also there is no obvious organization or sequence to its questions. The reasons for not answering various questions, however, are probably owing more to the nature of the questions than to the format. Questions about the future -- such as "what is your best guess that you will": graduate

with honors, change career choice, transfer to another college, find a job after college in the field for which you were trained, etc.? -- are generally skipped by 5% to 6% of the respondents. Questions about longrange aspirations or values are skipped by 5% to 8% of the students. Also, as noted earlier, questions about political and personal attitudes are typically skipped by 5% to 8% of the students. From one perspective, these are not large percentages; and the conclusions one draws from those who do respond would not be changed in any significant way if everyone had responded. From another perspective, these percentages of missing cases may represent the tip of a deeper and larger problem about the validity of students responses. There is no doubt that some students do not like some of the questions. During the time of student activism in the late 1960s, there were organized student protests against answering the sort of questions that are still included in todays edition. At the end of the questionnaire, 26% of the students do not give permission to include their ID number on any tape for future research or follow-up study. This undermines the validity of the data base for longitudinal studies. Moreover, when one realizes that the response rate to a mailed follow-up questionnaire may be only 50% or even less, then, together with the 26% refusal to be involved, one is left with a respondent population that may be only 1/3 or 1/4 of the population one should have.

Missing cases have also been noted for a third instrument -- Pace's College Student Experience questionnaire. Later in this report a detailed examination of the reliability and validity of responses will be presented. At this point, only the data about missing cases are reported. Most of the questionnaire consists of 142 college activities to which the

students respond by indicating whether and how often they have engaged in them during the current school year. These are, for the most part, quite specific events, and apparently quite easy to recall. Based on the responses of about 7,500 undergraduates, the number of missing responses was rarely more than 1%, and never more than 2%. These activities are grouped into scales, usually of 10 items, to which an activity score can be computed. If any item in a scale is not answered no score is computed. The number of missing cases in these scale scores is, in most scales, about 2% or less, and never more than 4%. In other parts of the questionnaire students are asked to indicate how much progress they believe they have made with respect to various goals or objectives, how well satisfied they are with college, and how they would characterize the college environment along various dimensions. The missing cases to these items are often fewer than 1% and never more than 2%. In another brief section of the questionnaire students are asked to indicate about how many textbooks they read, how many non-assigned books, how many essay exams they had, and how many other written reports they made during the current school year. The percent of missing responses was typically from less than 1% to 2%, except among students in not highly selective liberal arts colleges where there were 3% to 4% missing cases. No obvious explanation comes to mind for these somewhat larger percentages. With respect to the usual background items -- age, sex, year in school, etc. -- there are typically no more than 1% or 2% missing cases, except for the questions about the student's major field where the percent of missing cases ranges from 3% to 6% at different types of institutions. Unlike the ACT and HERI questionnaires which are given to beginning freshmen, the CSEQ is answered by undergraduates in

general, not just by freshmen, so that most of them do in fact have a definite major field. Why there should be from 3% to 6% omits is a mystery. Of course, not all possible major fields can be listed in the questionnaire so that students may wonder where to classify their own major. Also, especially in the more heterogeneous colleges, and also in the most selective ones, there may be more interdepartmental majors or other special options. Apparently, instead of checking "other" as the proper response, they just omit the item.

The College Student Experiences Questionnaire: A Brief Description.

To understand some of the analyses to be reported next, some knowledge about the content of this questionnaire may be helpful. The questionnaire is meant to be filled out by undergraduates toward the end of the academic year. It is an eight page, 8 1/2 by 11 format, with the cover page indicating what its all about, and stating that "we do not ask you to write your name anywhere in this questionnaire; but we do need to know where the reports come from, and that is why each questionnaire has a number on the back page -- certain blocks of numbers tell us that those questionnaires come from your college". The first 1 1/2 pages consist of "Background Information" -- the usual questions about age, sex, year in school, college residence, major field, parents education; and also time spent on academic work, time on a job, main source of funding for college, grades, race, and citizenship. The next 3 1/2 pages are labeled "College Activities". There are 142 activities, grouped into "scales" or topics labeled library experiences, experiences with faculty, course learning, art-music-theater,

student union, athletic and recreation facilities, clubs and organizations, experiences in writing, personal experiences, student acquaintances, science/technology, dormatory or fraternity/sorority, topics of conversation, and information in conversations. The directions are: "In your experience at this college during the current school year, about how often have you done each of the following?" The responses are "never", "occasionally", "often", and "very often". The activities are fairly specific so that the student would presumably recall accurately whether he had ever done them; but of course the frequency estimate is entirely a subjective response. Examples of activities are as follows: read something in the reserve book room or reference section, made an appointment to meet with a faculty member in his/her office, summarized major points and information in readings or notes, gone to an art gallery or art exhibit on the campus, meet your friends at the student union or student center, played on an intramural team, worked on a committee, asked other people to read something you wrote to see if it was clear to them, sought out a friend to help you with a personal problem, made friends with students from another country, practiced to improve your skill in using some laboratory equipment, gone out with other students for late night snacks, talked about current events in the news, referred to something a professor said about the (conversation) topic.

The next brief part of the questionnaire asks students to report how much reading and writing they have done, and how well satisfied they are with college.

The next main topic is labeled "The College Environment". This consists of eight rating scales on which students report their impressions

of the emphasis or stress there is in the environment on such aspects of students' development as academic and scholarly qualities, esthetic and creative qualities, being critical and analytical, vocational and occupational competence, and the general relevance and practical values of the courses; also their impressions of the personal relationships in the environment, ranging from supportive, helpful, considerate to alienated, unsympathetic, and rigid with respect to the relationships among students, between students and faculty, and with administrative personnel. Finally, the last section, labeled "Estimate of Gains", lists 21 goals or objectives of college education and asks students as follows: "In thinking over your experiences in college up to now, to what extent do you feel you have gained or made progress in each of the following respects?" The responses are "very little", "some", "quite a bit", and "very much". Here again, the responses are entirely subjective.

From one perspective, this College Student Experiences questionnaire includes features that some think should be avoided, if possible. The ratings are entirely subjective, the estimates of the amount of gain are entirely subjective, and the reports of frequency of participation in activities are entirely subjective. What follows next are examples of how subjective responses can be objectively validated.

Test-Retest Comparisons

In the absence of any major changes in the campus environment or facilities or admissions policy, one would expect some consistency in the amount, scope, and quality of effort revealed by students' responses to the

activities scales by different but comparable samples. Recently, several colleges have used the College Student Experiences questionnaire twice -- once in 1984, and again in 1985. Such comparisons are not, strictly speaking, an indication of the reliability of self-reports. The answers come from different students and from a different time. Some changes in the responses may reflect true changes, not random changes or errors of measurement. Nevertheless, if one found substantial variations in the responses of two different but similarly selected samples, even though a year apart, one would worry about the dependability of the results.

The best test-retest example comes from Denver University. It is best in the sense that the sample size was fairly large -- 635 in the spring of 1984 and 661 in the spring of 1985. The samples were selected in the same way, the response rate was similar, and the two goups did not differ in such population descriptors as age, sex, year in school, major field, grades, residence, transfers, etc.. No attempt is made to compare the responses to every item in the questionnaire. Rather, to get a general indication of stability, comparisons are made between the mean scores on each of the 14 activity scales, and the mean ratings on each of the environmental characteristics. It is not appropriate to report the scores on these matters, but it is permissible to report the differences between the 1984 scores and the 1985 scores. The second test-retest example comes from Case Western Reserve University -- with a sample of 779 students in the spring of 1984 and of 376 in the spring of 1985. The characteristics of the two samples are nearly identical with respect to age, sex, year in school, transfer status, major field, etc.. The third example comes from Keuka, a small college for women in upstate New York -- with 148 in the

1984 sample and 130 in the 1985 sample. The groups were similar in all respects except one: the 1985 sample had a larger proportion of freshmen.

On the 10-item activity scales the possible range of scores is 30 points; 36 points on the three 12-item scales; and 20 points on the one 6-item scale. The typical standard deviations are 5.7 on the 10-item scales, 6.0 on the 12-item scales, and 3.2 on the 6-item scale. A glance at the list of differences in the table quickly reveals that at all three schools the magnitude of differences is usually less than one point. This is true of 13 out of 14 scales at Denver, all 14 at Case Western Reserve, and 10 of the 14 scales at Keuka. In fact, at Denver the difference in mean scores between the 1984 and 1985 samples is .5 or less on 10 of the scales, at Case Western Reserve the differences are .5 or less on 13 of the 14 scales; and at Keuka on 6 of the 14 scales.

If comparable scores from comparable samples, even though a year apart, is an indication of test reliability, then obviously these student self-reports are very stable and dependable. At Denver, where there is a significant difference of 2.4 points on the Student Union scale, the explanation is a good example of a change in results owing to a change in conditions. During 1984 at Denver a new student union and activity center was under construction; 1985 was the first full year of its operation, and, not suprisingly, the activity score for students' use of the union increased significantly. At Keuka the differences between mean scores, although greater than 1.00 on four of the scales, are not statistically significant.

From these comparisons, one can surely conclude that self-reported activities and self-reported ratings of environmental characteristics are dependable and consistent.

Test-Retest Mean Differences -- 1984 and 1985
In Activity Scale Scores and Environment Ratings

Activity Scales	Denver Univ.	Case Western Reserve	Keuka College
Library Experiences	.3	.4	.7
Experiences with Faculty	.3	.1	1.3
Course Learning	.6	.2	0
Art, Music, Theater	.5	.4	.1
Student Union	2.4	.2	.4
Clubs and Organizations	.5	.9	.5
Athletic and Recreation	.1	.4	0
Experience in Writing	.5	.2	.2
Personal Experiences	0	.3	.7
Student Acquaintances	.6	.1	1.4
Science/Technology	0	.5	1.2
Dormatory or Fraternity/Sorority	.9	0	1.5
Topics of Conversation	.5	.4	.6
Information in Conversations	.1	.2	.8
Environment ratings			
Academic	.1	0	.1
Esthetic	.1	.2	.2
Critical/analytical	.1	.1	.1
Vocational	.2	0	.2
Personal Relevance	.2	.2	.1
Students	.3	0	.1
Faculty	1.3	0	.2
Administration	.1	0	.4

External Validity: Self-reported gains vs objectively known achievement

Over the past 50 years hundreds of thousands of college students have taken objective achievement tests in various college subjects, tests constructed by national testing agencies. Certain conclusions from all this testing are so consistent, and so obvious, that it almost seems unnecessary to state them; but if one is to document that self-reported achievement corresponds to objectively tested achievement, then some examples of the test evidence must be given. The examples that follow are reported in Pace, Measuring Outcomes of College, Jossey-Bass, 1979.

The first example shows the relationship between credit hours and test scores from the Pennsylvania study in 1928. The obvious conclusion is that students learn what they study, and the more they study the more they learn. Students who had the most credit hours in the natural sciences had the highest test scores on the natural sciences test items. The same is true for credits and scores in language, literature and fine arts, and also for credits and scores in social studies.

Credit Hours and Test Scores: 4500 Seniors from 45 Colleges in Pennsylvania, Tested in 1928

Natural Sciences Credits	Natural Sciences Test Scores
High: 55 or more	120
Statewide average: 37	78
Low: 6 or fewer	46

Language, Literature, and Fine Arts Credits	Language, Literature or Fine Arts Test Scores
High: 67 or more	250
Statewide average: 42	168
Low: 12 or fewer	111

Social Studies Credit Hours	Social Studies Test Scores
High: 97 or more	292
Statewide average: 52	241
Low: 12 or fewer	196

The second example, some forty years later, comes from the Area tests of ETS's Undergraduate Assessment Program. The test results are based on 47,000 seniors from 211 colleges in the years 1969-1971. For each of the three Area tests -- Humanities, Natural Sciences, and Social Sciences -- the average score for all seniors is compared with the average score of seniors whose "area of interest" corresponds to the subject matter of the test. The scores are standardized scores in which the standard deviation is 100 points. In the humanities area the differences between the two groups is 55 points. In the natural sciences area the differences are 57 points and 66 points. In the social sciences area the difference is 2 points. The sub-group is also part of the total group; and since 60% of the total group of seniors are also in the social science interest group, the difference is necessarily small.

UAP Area Tests: Approximately 47,000 Seniors from 211 Colleges in the Years 1969-1971

	Humanities Scores
All seniors	470
Seniors whose area of interest is in humanities (21% of all seniors)	525

	Natural Sciences Test Scores
All seniors	480
Seniors whose area of interest is in biological sciences (12% of all seniors)	537
Seniors whose area of interest is in physical sciences (7% of all seniors)	556

	Social Sciences Test Scores
All seniors	446
Seniors whose area of interest is in social sciences (60% of all seniors)	448

The third example comes from the UAP tests in designated major fields rather than from the more general Area tests. These results are shown in relation to the number of courses students had taken in their major -- fewer than eight vs eight or more courses. It is unlikely that , in one's major field, one would have fewer than six courses and still qualify as a major. Most likely, the comparisons are between students who have had 6 or 7 courses vs those who have had 8 to 12 courses. Again, the more one studies a subject the more one knows about it.

Given these obvious conclusions from decades of achievement testing, one can surely use them as external validity in relation to self-reported achievement. The College Student Experiences questionnaire, in the section labeled Estimate of Gains, lists 21 objectives. Students are asked "to what extent do you feel you have gained or made progress in each ...?" They could check "very little", "some", "quite a bit", or "very much". Not all of the objectives are associated with a specific major field, or even with any course-related experience -- objectives such as "ability to fuction as a team member", "ability to learn on your own, pursue ideas, and find information you need". There are, however, eight goals that are related to the curriculum, and specifically to a major field within the curriculum, or to a specific type of subject-matter. These subject-matter goals include Fine Arts, Literature, English (writing), Science, Technology, Computers, Quantitative thinking, and Philosophy/Cultures. If student self-reports are valid they should show the same results that test scores show -- higher achievement (progress) by students whose major field is similar to the objectives as compared with the average of all students -- and this is exactly what the results do, in fact, very clearly show.

Scale Scores of Seniors on Major Field Tests of the Undergraduate Assessment-Program, 1969-1971, Related to Number of Courses Taken in the Major Field

	Fewer than eight courses	Eight or more courses	Difference
Sciences and Engineering Tests			
Biology	539	566	+ 27
Chemistry	510	539	+ 29
Engineering	506	528	+ 22
Humanities Tests			
History	468	491	+ 23
Literature	455	491	+ 36
Philosophy	514	551	+ 37
French	448	486	+ 38

Note: The number of students tested varies by major field, ranging from approximately 1,000 to 8,000.

The data presented here are from a composite of 13,650 undergraduates from 49 colleges and universities who responded to the CSEQ in the spring of 1983, 1984 or 1985. Only those colleges that had given the questionnaire to all four undergraduate classes are included in these composite results. Note also that knowledge or progress is necessarily less among freshmen or sophmores who have not yet accumulated many credits in what is or will be their major field, than it would be among juniors and seniors who, by definition, have accumulated a much larger number of credits in their chosen major field. For some, then, the "major" may reflect an "area of interest" and for others it may be a course program nearly completed.

In the list below, the first four goals are related to the subject matter of arts and humanities, and the second four goals are related to the sciences. Among students who identified their major field as "Arts (art music, theater, etc.)", 92% reported substantial gain ("quite a bit" plus "very much") toward the objective "developing an understanding and enjoyment of art, music and drama". This high percentage contrasts with 29% among students in general. For the objective related to literature, 74% of humanities majors reported substantial gain compared with 33% for students in general. With respect to writing clearly and effectively, 80% of the humanities majors reported substantial progress compared with 57% of students in general. The goal described as "becoming aware of different philosophies, cultures, and ways of life" is not so clearly tied to classroom subject matter in the sense that students' interpersonal

experiences might well contribute significantly toward its attainment; but presumably courses in philosophy, history, anthropology, etc. would also be influential. The results show substantial progress reported by 70% of humanities majors, and 64% of social sciences majors, compared with 51% by students in general.

Comparisons of Self-Reported Gains with Known Data About Achievement

Gains in developing an understanding and enjoyment of art, music, and drama

ARTS majors reporting substantial gains	92%
average of all students	29%

Gains in broadening your acquaintance and enjoyment of literature

HUMANITIES majors reporting substantial gains	74%
average of all students	38%

Gains in writing clearly and effectively

HUMANITIES majors reporting substantial gians	80%
average of all students	57%

Gains in becoming aware of different philosophies and cultures

HUMANITIES majors reporting substantial gains	70%
SOCIAL SCIENCE majors reporting substantial gains	64%
averge of all students	51%

Gains in understanding the nature of science and experimentation

BIOLOGICAL SCIENCES majors reporting substantial gains	85%
PHYSICAL SCIENCES majors reporting substantial gains	76%
average of all students	36%

Gains in understanding new scientific and technical developments

BIOLOGICAL SCIENCES majors reporting substantial gains	74%
PHYSICAL SCIENCES majors reporting substantial gains	66%
ENGINEERING majors reporting substantial gains	66%
average of all students	31%

Gains in acquiring familiarity with the use of computers

COMPUTER SCIENCE majors reporting substantial gains	90%
ENGINEERING majors reporting substantial gains	65%
average of all students	32%

Gains in quantitative thinking -- understanding probabilities, proportions, etc.

PHYSICAL SCIENCES majors reporting substantial gains	68%
ENGINEERING majors reporting substantial gains	68%
average of all students	47%

In "understanding the nature of science and experimentation", substantial progress is claimed by 85% of biological sciences majors and 76% of physical sciences majors, compared with 36% for students in general. A similar result is shown for "understanding new scientific and technical developments", with percentages of 74% and 66% for scientific and technical majors, compared with 31% for the average of all students. The contrasting percentages for the goal "acquiring familiarity with the use of computers" are even sharper -- 90% of majors in computer science indicating substantial progress compared with 32% for the average of all students. With respect to quantitative thinking, students majoring in fields where much quantitative thinking is required -- engineering, and physical sciences -- are most likely to claim substantial progress (68%) compared with 47% among students in general.

All of the above results document the external validity of students self-reports. When asked to rate their progress toward goals that are obviously related to the subject matter of college courses, the ratings are totally congruent with what we know from achievement test scores and from the relationship between credit hours or amount of study and measued achievement.

One does not know the actual level of measured achievement (standardized test scores) that is associated with the students' self-estimate of gain. No doubt some students who rate their own progress as "quite a bit" may have higher achievement test scores than students at another college who rate their progress as "very much". Such discrepancies probably reflect institutional differences in academic selectivity and academic demands. The same variability applies to credit hours <u>vs</u> test

scores. While it is true that the more courses one takes in a subject the more one is likely to know about it, it is also true that some students who have taken 5 or 6 courses may get higher test scores than some students who have taken 9 or 10 courses. But the averages are consistent. Sorting students according to course work (major field) or according to achievement test scores (major field) or according to self-reported progress (in major fields) all produce the same conclusions.

Internal Reliability: Consistency in responses to similar items

In the Science/Technology activity scale there are three activities that clearly involve conversation about science. These items, together with the percent of students who said they engaged in them frequently, are shown below:

Science activities	% Frequently among Bio.Sci. majors	Phys.Sci. majors	Engr. majors	Average of all students
Tested your understanding of some scientific principle by seeing if you could explain it to another student.	70	69	69	34
Showed a classmate how to use a piece of scientific equipment	43	35	34	18
Attempted to explain an experimental procedure to student	43	42	41	15
Conversation topic				
Science -- theories, experiments, methods	57	53	58	21

The conversation item appears in a different part of the questionnaire. Presumably, the percent of students who say they have frequently talked about science with other students should have some similarity to the percent who said they had tried to explain a principle, a procedure, and the use of equipment to another student. The responses were, in fact, very similar.

A similar comparison can be made in the arts. In the activity scale labeled Art, Music, Theater there are three "talk about" items, and later, among the conversation topics there is a topic described as "Fine arts - painting, theatrical productions, ballet, symphony, etc.". Here are the results.

Art, Music, Theater activities	% Frequently among Arts majors	Among all students
Talked about art (painting, sculpture, architecture, artists, etc.) with other students at the college	68	17
Talked about music (classical, popular, musicians, etc.) with other students at the college	73	35
Talked about the theater (plays, musicals, dance, etc.) with other students at the college	58	20
Conversation topic		
Fine arts -- painting, theatrical productions, ballet, symphony, etc.	78	17

Similar but not identical questions produce similar but not identical answers. The general congruence shown above can be regarded as an indication of internal reliability.

Internal Validity: finding plausible connections

For the attainment of many goals of higher education there is no readily available objective documentation and in some cases no external evidence at all. One can use tests and credits when the goals are related to the curriculum or to particular courses and major fields. But what does one use for an external criterion when the goals are self-understanding, understanding others, good health habits, functioning as a team member, etc.?

In this part of the report several examples of internal consistencies that should be found are used to make the case for the credibility of self-reports. The first example is surely a connection that should exist. The activities -- setting performance goals, following a regular exercise schedule, and keeping a record of progress -- are, to a considerable extent, behavioral indicators of what is involved in "developing good health habits and physical fitness". The tabulations show that students who report "very much" progress toward this goal are much more likely to set goals, follow a schedule, and keep a record than students whose self-rated progress is lower.

Similar tabulations are shown for several other goals. In every case, the behavior that surely should contribute to students' estimated progress is clearly related to that progress. The differences in percents between "very much" and "very little" are uniformly large, the one being from 2 to more than 6 times larger than the other.

If student responses to the gains items or to the activity items were capricious or unreliable or invalid, the congruent and plausible

connections shown in the tables below would not occur. If what should be true is also true empirically, the credibility of self-reports is further documented.

Goal: Developing good health habits and physical fitness

Activity	Percent engaging in activity frequently among students who rate their progress as: Very Much	Quite a bit	Some	Very Little	Average of all students
Set goals for your performance in some skill (athletic)	77	58	36	23	45
Followed a regular schedule of exercise, or practice in some sport, on campus	71	53	28	14	38
Kept a chart or record of your progress in some skill or athletic activity.	28	15	6	3	11

Goal: Ability to function as a team member

Activity	Percent engaging in activity frequently among students who rate their progress as: Very Much	Quite a bit	Some	Very Little	Average of all students
Used outdoor recreational spaces for casual and informal group sports	40	27	15	7	23
Used facilities in the gym for playing sports that require more than one person	42	30	18	10	26
Played on an intramural team	36	26	15	7	22

Goal: Understanding yourself -- your abilities, interests, and personality

Activity	Percent engaging in activity frequently among students who rate their progress as: Very Much	Quite a bit	Some	Very Little	Average of all students
Read articles or books about personal adjustment and personality development	38	25	20	15	28
Asked a friend to tell you what he/she really thought about you	33	21	14	12	23
Identified with a character in a book or movie and wondered what you might have done under similar circumstances	56	44	36	32	46

Goal: Understanding other people and the abiltiy to get along with different kinds of people

Activity	Percent engaging in activity frequently among students who rate their progress as: Very Much	Quite a bit	Some	Very Little	Average of all students
Made friends with students whose interests were very different from yours	73	57	38	32	59
Made friends with students whose family background (economic and social) was very different from yours	78	63	44	36	63
Had serious discussions with students whose political opinions were very different from yours	45	33	26	22	35

Goal: Becoming aware of different philosophies, cultues, and ways of life

Activity	Percent engaging in activity frequently among students who rate their progress as: Very Much	Quite a bit	Some	Very Little	Average of all students
Made friends with students whose race was different from yours	62	50	40	33	46
Made friends with students from another country	50	24	24	20	31
Had serious discussions with students whose philosophy of life or personal values were very different from yours	64	48	33	25	43
Had serious discussions with students whose religious beliefs were very diffrent from yours	55	40	28	22	36
Had serious discussions with students from a country different from yours	42	25	15	13	23

Summation

Claims for the credibility of student self-reports can be supported by:

1. Evidence of test-retest consistency.

2. Congruence with externally known facts, when such facts are available.

3. Similar answers to questionnaire items, when questions are asked in more than one way.

4. Congruent, or expected, connections between items that presumably should have connected responses -- as for example between behavior and progress.

One final note may be important. Some psychometricians and survey research analysts point out that the context within which questions are asked may influence the response. In the College Student Experiences questionnaire, some people might claim that the answer to the Estimates of Gains items might be "contaminated" by all the preceding items. The gains might be reported differently if they were asked separately, or without the prior context in the questionnaire. There is, however, a very different way of regarding this matter. If the gains items were presented alone, without any context, the responses would be all the more influenced by personal idiosyncrasies, and hence all the more likely to produce random variations. By putting the gains items at the end of the questionnaire, one increases the credibility of answers. Everyone comes to these items with the same background, having recalled one's behavior during the year, having characterized the college environment, having reported how much one has studied, what grades one has received, etc. so that, for everyone, the

estimate of gains becomes a more or less commonly based and thoughtful summary of the college experience, and therefore has a greater reliability.

Finally, as a capstone illustration of what can be done to assess the reliability of self-reports, we apply some multivariate statistical procedures which bear upon the predictive and construct validity of certain parts of the College Student Experiences questionnaire.

Multivariate statistical procedures

In this part of the report we describe the use of common multivariate statistical procedures to assess the validity of self report. The goal is to demonstrate that for surveys that allow internal validity checks, one can go beyond item-by-item validity to assessing the validity of self report at the construct level. These techniques are applied to a sample of 6,000 undergraduates who provided responses to the College Student Experiences Questionnaire (CSEQ).

Three techniques were applied to two types of scales and one background variable of the CSEQ. The background variable is academic major coded as: 1) Arts; 2) Biological Sciences; 3) Business; 4) Computer Science; 5) Education; 6) Engineering; 7) Health related fields; 8) Humanities; 9) Physical Sciences; 10) Social Sciences. The two types of scales are composed of 13 subscales from the Quality of Effort (QE) measures, and 21 items from the Estimate of Gains (EG) measures.

The first statistical procedure is discriminant analysis with special attention paid to the classification phase of the analysis. The discriminating variables are the EG items while the classification variable

is academic major. Since the number of undergraduates in each major are not the same, special a priori weighting is given to the samples during the classification phase. The rationale for using discriminant analysis in this context is that those who major in certain academic disciplines probably make the most gains in those areas related to that discipline. Hence, if one knows a student's set of responses to the gains items, one should be able to predict that individual's major. To the extent that this is true, it might be argued that the EG measures provide valid self report measures of gains.

The second procedure is canonical correlation analysis applied to the QE subscales and the EG items. This procedure attempts to find a set of linear combinations (canonical variates) within a scale that are maximally correlated with linear combinations formed from the other scale. To the extent that these canonical variates are interpretable, we would expect high canonical correlations among those variates from each set that have something in common. Often it is the case that canonical analysis obscures the simple factor structure that might exist within a set of items. To address this problem, the third procedure is to factor analyze the QE subscales and EG items separately, rotate the factors for maximum interpretabilty, calculate factor scores, and correlate factor scores using simple Pearson correlations. It is expected that Pearson correlations should show high correlations among those factors that are substantively related.

Discriminant and classification analysis were performed using ten academic majors and twenty EG items. The EG items were chosen to correspond as closely as possible to the academic majors, hence the item

related to gains in vocational training was omitted since no major was uniquely vocational.

The results of the classification phase are displayed in the following table. The table shows the percentage of those who were classified into their known majors on the basis of the discriminant analysis. The diagnonal represents the percentage of correct classifications, while the off-diagonal represents the misclassifications. It can be seen that the EG responses tend to do well in predicting academic major. For example, 61% of all art majors were correctly classified as being art majors on the basis of the discriminant analysis. Certain incorrect classifications did occur; but the misclassifications were in a sensible direction. For example, physical science majors (including chemistry and math) were more often classified as biological science majors (including biochemistry) or engineering majors. Overall, these results lend support to the claim that self report of gains as measured by the EG data are valid in that they adequately predict a relatively objective measure of academic field where most gains should occur.

The results of the canonical analysis are displayed in the next table. Here, only the first two canonical variates extracted from each set of measures are presented. Note that the standardized canonical coefficients can be loosely interpreted as factor loadings.

Inspection of the standardized canonical coefficients for the QE subscales suggests that the first canonical variate is dominated by the QE subscale related to Science and Technology. The first canonical variate to the EG data appears to be dominated by those items related to computer knowlege and Science/Technology. The squared canonical correlation between

Classification Analysis of Academic Major
on Basis of Discriminant Analysis*

	Predicted Major										
True Major	Art	Bio	Bus	C/S	Ed	Engr	Heal	Hum	PhyS	SocS	Total
Arts	61	4	11	1	0	0	0	11	0	11	100
Bio	1	52	7	1	0	15	13	0	0	10	100
Bus	2	1	73	4	2	3	2	1	0	12	100
CompSci	1	1	33	45	0	9	0	1	0	1	100
Educ	9	2	34	2	15	3	9	7	0	18	100
Engr	1	10	14	8	0	57	4	0	1	5	100
Health	2	20	15	0	4	4	34	1	0	19	100
Human	8	3	11	1	3	1	3	38	0	32	100
Phs/Sci	0	33	17	7	1	25	5	1	3	7	100
Soc/Sci	4	7	28	2	4	3	4	9	0	39	100
TOTAL %	5	11	33	6	3	11	7	6	0	17	100

* Entries are in percentages.

Standard Canonical Variates for
QE Scale and EG Scale Items

EQ Canonical Variates			EG Canonical Variates		
Subscales	QE1	QE2	Items	EG1	EG2
Library	.02	.05	Professional Sci or Scholarly	.03	.10
Faculty	- .06	.10	General Education	- .02	- .02
Course Work	- .05	.15	Career Development	- .07	.03
Art, Music Drama	- .16	.66	Art, Music, Drama	- .14	.64
Student Union	.00	- .05	Literature	- .07	.16
Recreation	- .00	- .05	Writing	- .17	- .13
Clubs	.00	.04	Computers	.51	.09
Writing	- .14	- .03	Philosophies/ Culture	- .08	.09
Personal Experiences	- .14	- .03	Ethical Standards	- .02	.08
Acquaintances	- .04	.01	Personality	- .03	.01
Sci/Tech.	.94	.21	Understanding People	- .05	.08
Conv. Topics	.03	.21	Team Work	.01	.04
Information	- .06	.20	Physical Fitness	- .03	- .06
			Science Experim.	.27	.01
			Science/Technology	.32	.15
			Technology/Hazards	.00	.14
			Analytical Thinking	.05	- .01
			Quantitative Thinking	.16	- .15
			Similarities and Differences	- .11	.20

these two canonical variates is statistically significant [R^2 = 0.61, F (260, 44745) = 42.979, $p < .000$].

Inspection of the second canonical variates for both sets of measures reveals a similar consistent picture. The second canonical variate for the QE scale is related to art, music, and theater, while the second canonical variate of the EG responses is related to gains in understanding art, music, and drama. Again, the squared canonical correlation between this pair of variates is statistically significant [R^2 = 0.38, F (228, 41681) = 29.011, $p < .000$]. Subsequent canonical variates were difficult to interpret.

It can be seen that the canonical analysis gives a useful, though limited, picture of the internal validity of the two self report measures. Again, it should be noted that this procedure examined validity of self report at the construct level, where the canonical variates can be taken as representing the constructs, though perhaps not in the factor analytic sense.

On the basis of previous research, four factors of the QE scale and five factors of the EG items were independently extracted and obliquely rotated to simple structure. The four factors of the QE scale were labeled 1) Personal/Social; 2) Academic/Intellectual; 3) Clubs/Organizations; 4) Science. The five EG factors were labeled 1) Personal/Social; 2) Science/Technology; 3) General Education; 4) Intellectual; 5) Vocational. A matrix of Pearson correlations among the factor scores obtained from the factor analysis is displayed in the next table. Although most of the correlations are large and significant, those that are highest are among

Intercorrelations Among Factor Scores

Estimated Gains Factors	Quality of Effort Factors: P/S	A/I	C/O	Sci
P/S	.50	.42	.44	.11
S/T	.19	.13	.04	.62
G/E	.42	.45	.31	.07
Intel	.36	.36	.22	.43
Voc	.29	.29	.24	.25

those factors that have something in common. For example, the gains in personal and social development factor is most highly correlated with QE factor measuring personal and social aspects such as student acquaintences, personal experiences, and topics of conversation.

Two points can be made with regard to the above analyses. First, the application of multivariate statistical procedures for assessing broad construct validity of self report has potential. It should be pointed out however, that construct validty in the factor analytic sense was only explained via the factor score correlations. Secondly, with respect to the CSEQ, and the QE scales and EG items in particular, evidence does exist for claiming a certain degree of validity in these self report measures. The result of all three analyses present a picture of a questionaire that is consistent with respect to self report predictive validity and self report construct validity.

CONCLUDING COMMENTS

This report is obviously not a definitive document about the credibility of questionnaire survey responses by college students. It has aimed, nevertheless, to show that there are many ways to confirm the accuracy, reliability, and validity of student self-reports. It has also noted, from examples in higher education, and from examples in the larger area of public opinion polls and other general surveys, some of the common sources of measurement errors and errors of substance.

In academic surveys the high proportion of students who do not reply to the questionnaires they have received is a most serious problem. One wonders whether rigorous follow-up messages would make a big difference, or whether the magnitude of the non-respondent problem reflects a deeper rejection of such inquiries. Twenty years ago one could expect about two thirds of college students to respond to a questionnaire. Today, one is grateful if 50% respond. Times change. Nearly 50 years ago, in a study I directed of former university students, including some who had graduated and some who had not, we got returns from 70% of those who received the questionnaire. The questionnaire was 52 pages long and took about two hours to answer. But that was before the invention of television! (Pace, They Went to College, University of Minesota Press, 1941).

My own belief is that the likelihood of good returns is enhanced by the recipients' opinions about the importance of the topic, its perceived relevance to one's experience, one's regard for the source of the inquiry and the likely use or value of the results, the clarity of questions and the ease or confidence one has in answering them, and the overall

attractiveness of the design, format, typography, etc. of the instrument. I also believe that unless these conditions are reasonably well met, even vigorous follow-up efforts will have little influence on the response rate, and even when some increase in response rate is achieved I would be skeptical about the integrity of those added responses.

Perhaps the second most common weakness in questionnaires by academic organizations is the inclusion of questions that are quite likely to have unreliable or invalid answers. These may be questions about vague concepts, questions about topics that students have not previously thought about, questions about values or life goals or future plans. Similar weaknesses are evident in public opinion polls that ask for opinions about ambiguous or undefined concepts such as national defense, foreign aid, national health, etc. The unfortunate consequence is that pollsters and public alike think that the results reflect public attitudes toward the matter, when in fact the topic is complex, can be phrased in a variety of proper ways, and all one has done is to tally answers to the particular question which is not well or uniformly interpreted in the first place. Questions about future expectations can be very clear -- for example, "Do you expect to have any (more) children?" But it is difficult to know just what is being measured or revealed by answers to questions that different people can interpret in different ways.

A final issue is the use of single questions versus the use of scales or combinations of questions that can be added together to produce a score or index. Commercial agencies rely on single items. Scale development is complex, time-consuming, and costly; and for public opinion polling agencies the presumed benefit is not worth the price. A scale is not

always better (more reliable and valid) than a single item. In most academic surveys, however, the topics of inquiry tend to be rather global rather than narrowly explicit. In these cases there is merit in thinking about questionnaire construction in ways somewhat similar to thinking about test construction.

Whatever the topic of inquiry, it may well be that one of the most important elements to consider in writing the questions is the nature of judgment required to answer them. If the judgment or thought process is one of recall, is the thing or condition to be recalled clear and are the respondents able to recall accurately? If the judgment is one of comparison, is the base for the comparison clear and do the respondents have the experience or knowledge needed to make the comparison with reasonable confidence. If the judgment or thought process to answer the questions is one of generalizing or inferring, do the respondents understand what is to be generalized? Many survey questions would probably yield better answers if the writers always asked themselves such questions as: Does the respondent have the knowledge or experience to give a useful answer? Will different people interpret the quesiton in the same way? Will the answer be accurate? What can I conclude or interpret from the answers to this question?

The quality of questionnaire answers (reliability, validity, credibility) depends most of all on the quality of the questions.